

# **STATISTICAL ANALYSIS OF VEHICLE TEST DATA FROM A CROSSOVER STUDY**

**April 13, 2001**

## ***Prepared for***

United States Environmental Protection Agency  
National Vehicle and Fuel Emissions Laboratory  
Assessment and Standards Division  
2000 Traverwood Drive  
Ann Arbor,  
Michigan, 48105

## ***Prepared by***

Jonathan Cohen  
ICF Consulting  
101 Lucas Valley Road, Suite 230  
San Rafael  
California 94903

## BACKGROUND AND EXPERIMENTAL DESIGN

A large corporation recently ran a sizable test program to determine the emission effects of introducing certain gasoline components. These fuel effects require mileage accumulation before their impact on emissions and fuel economy becomes pronounced. There is some evidence of carry-over so that the impact of the fuel component may carry over for some driving time after use of the fuel component is discontinued. The fuel component effect may require significant mileage accumulation using the reference fuel before dissipating.

The test program was carried out on 28 vehicles, four vehicles of each of seven types. The seven types were primarily chosen to represent different sets of emissions certification standards. A crossover design was used. After selection, each vehicle was driven for 1,000 miles on the reference fuel and then the emissions and fuel economy were tested. The vehicle was driven for another 8,000 miles using a specified fueling regime, followed by a second set of emissions and fuel economy tests. The vehicle was driven for an additional 8,000 miles using a different specified fueling regime, followed by a third set of emissions and fuel economy tests. Possible fueling regimes are:

- T: Test fuel used continuously
- R: Reference fuel used continuously
- A: Reference fuel and test fuel used alternately for each tank fill, starting with the reference fuel.

In the following we shall refer to R as the reference fuel, T and A as the test fuels, T as the continuous test fuel, and A as the alternating test fuel.

For each vehicle type the four vehicles were allocated to one of following four fueling regime test sequences for the two phases of 8,000 miles of mileage accumulation:

- Vehicle 1: R, T
- Vehicle 2: T, R
- Vehicle 3: R, A
- Vehicle 4: A, R

The main purpose of the study was to evaluate the direct and carryover effects of the test fuels A and T on NO<sub>x</sub> emissions measured over the Federal Test Procedure (FTP). The direct effect measures the change in emissions while the test fuel is being used. The carryover effect measures the change in emissions after the test fuel has been used and replaced by the reference fuel. More detailed definitions are given below. The following statistical analysis focuses on developing models for FTP NO<sub>x</sub> emissions, but the final set of models were also applied to FTP CO and HC emissions.

## SUMMARY OF STATISTICAL MODELS

Numerous statistical models could be fitted to these data. Our approach was based on the following general formulation:

$$\log \text{NOx (vehtype, vehicle, miles)} = \mu + \lambda \times \text{miles} + \alpha \times \text{direct} + \beta \times \text{carryover} + \text{error}$$

In this general structure,  $\mu$  and  $\lambda$  are the intercept and slope, defining the vehicle baseline (at the beginning of the study) and mileage deterioration rates for the reference fuel, and the direct and carryover terms represent the fuel effects. More specifically,

NOx	=	Mean FTP NOx emissions measured at 0, 8,000 or 16,000 miles after the initial 1,000 miles accumulated on the reference fuel;
log	=	natural logarithm;
vehtype	=	vehicle type (a classification variable with seven levels);
vehicle	=	vehicle number (1, 2, 3 or 4, treated as a classification variable);
miles	=	thousands of miles after initial 1,000 mile reference fuel accumulation (a continuous variable with values 0, 8, or 16),
	=	phase number $\times$ 8;
$\mu$	=	intercept, possibly depending on the vehtype or vehicle;
$\lambda$	=	slope, possibly depending on the vehtype or vehicle;
direct	=	indicator (dummy variable) for the main effect of the test fuel compared to the reference fuel,
	=	1 if the test fuel was used for the last 8,000 miles 0 if the reference fuel was used for the last 8,000 miles 0 otherwise;
$\alpha$	=	direct effect of the test fuel, possibly depending on the vehtype, vehicle, type of test fuel (continuous or alternating), or miles,
	=	increment in expected log(NOx) attributable to the test fuel;
carryover	=	indicator (dummy variable) for the “engineering” carryover effect of the test fuel compared to the reference fuel,

	=	1 for phase 2, vehicles 2 and 4 (after sequences T, R or A, R) 0 otherwise;
$\beta$	=	“engineering” carryover effect of the test fuel, possibly depending on the vehtype, vehicle, or type of test fuel (continuous or alternating),
	=	difference between expected log(NOx) at the end of phase 2 for the fuel sequence Test fuel, R compared to the fuel sequence R, R;
error	=	random error (mean zero), depends on vehtype, vehicle, and miles. These errors may be correlated.

The coefficients  $\mu$ ,  $\lambda$ ,  $\alpha$ , and  $\beta$  may not be constant in the general formulation, but instead may depend upon other variables such as the vehtype or vehicle. Equivalently, there may be interactions between the intercept, slope or fuel effects and other variables. For example, if  $\lambda$  depends on the vehtype, then the model will include an overall slope (miles term) and interactions between vehtype and miles.

The “engineering” definition of carryover is constructed from the model to compare the emissions for the vehicle 2 and 4 test sequences with the emissions that would be expected to have accumulated on the reference fuel alone over the same 16,000 miles. The latter test sequence was not included in the experimental design. An equivalent mathematical formulation redefines carryover as

$\beta - \alpha(\text{phase 2})$	=	$\beta - \alpha$ if the direct effect is the same for both phases
	=	“statistical” carryover effect of the test fuel, possibly depending on the vehtype, vehicle, or type of test fuel (continuous or alternating),
	=	difference between expected log(NOx) at the end of phase 2 for the fuel sequence Test fuel, R compared to the fuel sequence R, Test fuel.

This is referred to as “statistical” carryover since it is the usual statistical definition of carryover for a two stage crossover design: the difference between the total responses (sums of phase increments) when the two treatments are given in the opposite order.

Crude estimates of the direct and carryover fuel effects can be developed from the averages across vehtypes and vehicles of the changes in log(NOx) over the two 8,000 mile phases. After removing the phase 0 outliers for the F150 vehicle 4 and for the LeSabre vehicle 1, these averages were:

Fuel	Phase	Mean change in log(NO <sub>x</sub> )
R (reference)	1	0.13
R	2	0.03
T (continuous)	1	0.09
T	2	0.10
A (alternating)	1	0.05
A	2	0.17

Averaging across the two test fuels and across the two phases gives an average increment of 0.10, which is 0.03 lower than the reference fuel phase 1 increment (0.13). Thus the estimated direct effect of the test fuels is  $\alpha = -0.03$ . (This calculation assumes that the underlying direct effects are the same for both test fuels and for both phases). The difference between the reference fuel increments for phases 1 and 2 is  $0.03 - 0.13 = -0.10$ . This difference is an estimate of  $\beta - \alpha$ , the “statistical” carryover, because the experimental design sequences either preceded or followed the test fuel by the reference fuel. The estimated “engineering” carryover,  $\beta$ , is therefore  $-0.10 + -0.03 = -0.13$ .

Note that the same approach applied only to the phase 1 data shows a stronger direct effect of -0.06. The validity of also using data from phase 2 in the analysis, despite the crossover design, is discussed below.

The more complex statistical models developed below lead to similar estimated fuel effects, as shown in Table 1. The two alternative model structures were developed from the NO<sub>x</sub> data, but have also been applied to the HC and CO data for comparison (see Table 3 below). For these two statistical models, the intercept and slope depend on the vehicle type and vehicle, and the direct and carryover effects are assumed to be constant (the same value for all vehtypes, all vehicles, both test fuels, and both phases). Both these models both show that for NO<sub>x</sub>, the direct effects are not statistically significant (at the five percent level), but the “engineering” and “statistical” carryover effects are much larger and are statistically significant in three of the four cases.

The finding of large carryover effects but small direct effects is counterintuitive, but there are several possible explanations. One possibility might be that vehicles 2 and 4 had some fundamentally different characteristics compared to vehicles 1 and 3, which would certainly impact the estimated fuel effects. Such fundamental characteristics might include mileage or fuel differences prior to the study. An examination of the initial odometer mileages given in Table 2 does not show a consistent pattern of numerical mileage differences although the initial mileages

for vehicle 1 were lower than those for vehicle 2 in five of the seven vehetypes and the initial mileages for vehicle 3 were lower than those for vehicle 4 in five of the seven vehetypes. Vehicle 1 minus vehicle 2 mileage differences ranged from -32K (C1500) and -22K (Caravan) to +41K (LeSabre). Vehicle 3 minus vehicle 4 differences were all between -8K (Accord, Escort) and +7K (F150) except for the +25K difference between Caravan vehicles 3 and 4.

It is also possible that the fuels used in vehicles 1 and 3 prior to the study were quite different to the fuels used in vehicles 2 and 4 (i.e., the fuels used prior to the study and the reference fuel used in phase 1). The initial 1,000 mile preconditioning on the reference fuel may have been insufficient to bring all four vehicles to a common starting point, so that the phase 1 differences may reflect inherent differences between the vehicles but the phase 2 data would be much less affected by these inherent differences due to their total 9,000 miles of preconditioning. Another physical explanation is the possibility that the full effect of the fuel additive is not realized until after 8,000 miles, so that the direct effects over 8,000 miles do not reflect the full impact of the additive and the carryover effects reflect the true impact. A further test program with different test and reference fuel mileage accumulations and more frequent FTP tests might resolve these issues.

**Table 1. Final models for estimating emissions impacts (changes in the natural logarithm of emissions) attributable to the fuel additive.**

Pollutant	Model	Direct effect	Standard error of direct effect	“Statistical” carryover effect	Standard error of “statistical” carryover effect	“Engineering” carryover effect	Standard error of “engineering” carryover effect
NO <sub>x</sub>	1	-0.022 (0.59)	0.042	<u>-0.112</u> ( $< 0.01$ )	0.037	<u>-0.135</u> (0.05)	0.066
	2	-0.025 (0.46)	0.034	<u>-0.072</u> (0.02)	0.030	-0.097 (0.08)	0.054
HC	1	0.045 (0.11)	0.027	<u>0.091</u> ( $< 0.01$ )	0.025	<u>0.136</u> ( $< 0.01$ )	0.044
	2	<u>0.063</u> ( $< 0.01$ )	0.022	<u>0.091</u> ( $< 0.01$ )	0.020	<u>0.155</u> ( $< 0.01$ )	0.036
CO	1	0.043 (0.30)	0.041	<u>0.127</u> ( $< 0.01$ )	0.037	<u>0.167</u> (0.01)	0.065
	2	-0.008 (0.76)	0.028	<u>0.096</u> ( $< 0.01$ )	0.025	0.088 (0.05)	0.044

Notes:

- (1) Model 1 has equal error variances for all vehicle types. Model 2 has different error variances for each vehicle type.
- (2) The FTP mean outliers for phase 0, vehtype F150, vehicle 4 and phase 0, vehtype LeSabre, vehicle 0 have been deleted.
- (3) Statistically significant effects at the five percent level are underlined. P-values (significance levels) are shown in parentheses.

**Table 2. Initial Mileage Accumulations (000's of miles)**

<b>Vehicle Type</b>	<b>Vehicle 1</b>	<b>Vehicle 2</b>	<b>Vehicle 3</b>	<b>Vehicle 4</b>
Explorer	23	24	21	24
C1500	25	57	26	33
Accord	30	26	18	26
F150	40	42	61	54
Escort	64	72	71	79
Caravan	75	97	106	81
LeSabre	88	47	63	67

## **MODEL DEVELOPMENT**

The model was developed using an iterative process. This process defines the outliers, the significant interactions, and the random and fixed effects in the final models. An initial outlier screening was applied to the triplet measurements before computing the phase averages. A general linear model (GLM, also known as analysis of variance) approach was then used to define the reference fuel intercept and slope terms, and the phase mean outliers were identified. After reformulating the GLM model as a mixed model (with fixed and random effects), the possible interactions between fuel effects and vehtype or miles were evaluated. Next, a variety of models were fitted to evaluate possible variance homogeneity and to model the effects of within-vehicle correlations (repeated measures). Finally, the differences between the two test fuels were tested to evaluate the possibility of equal test fuel effects. The same final model structures were fit to the NO<sub>x</sub>, HC, and CO data. The main results, with the two outlier phase means deleted, were compared with other outlier treatments.

### **Phase Means**

The first issue was to decide between basing the analysis on the individual tests or trials (either 2 or 3 FTP tests for each combination of vehtype, vehicle, and phase) or on using phase means (i.e., averages across triplets or pairs of trials).

The extra "third" test was applied if the difference between two of the emissions tests (for a given vehicle and phase) exceeded 35 % for HC, 70 % for CO, or 29 % for NO<sub>x</sub>. If all tests are used, then a disadvantage is that vehicles are given different weights in the analysis and that the vehicles with greater variability (leading to the need for a third test) will be the vehicles given

greater weight. Averaging across tests at each testing point (vehicle/phase combination) gives each vehicle equal weight but incorrectly assigns equal variability to arithmetic means of two and three tests. The averaging at each testing point will also reduce the number of degrees of freedom for estimating the model parameters, but this reduction will primarily impact the estimated error variance parameter rather than the fuel effects parameters.

The following options can be considered:

1. Use all tests.
2. Use the arithmetic mean of all emissions tests at each testing point. This removes the problem of vehicles with third tests getting extra weight in the analysis and also lessens the impact of aberrant values among the three emissions tests. However, if the three tests have the same underlying test-to-test (error) variance, then the means from three tests will have smaller standard errors, and the statistical model will not take this into account.
3. Only use the first two emissions tests at each testing point. This removes the problem of vehicles with third tests getting extra weight in the analysis but does not deal with potential outliers among the first two emissions tests. The fact that a third test was necessary suggests that one or other of those two tests was an outlier.
4. Apply an outlier test to the triplets of emissions tests, and delete any test found to be an outlier. A suitable test is the Hawkins-Perold test as used in the Auto/Oil Program: For each triplet, the maximum deviation from the sample mean is divided by the pooled standard deviation estimate (computed from all triplets and pairs of emissions tests) and compared to the tabulated critical value. These tables are given in the report *Auto/Oil Air Quality Improvement Research Program, Description of Phase II Working Data Set* (Cohen and Noda, 1995). Tests with deviations above the critical value are determined to be outliers. This procedure deals with potential outliers among the emissions tests, but may not remove the problem of vehicles with third tests getting extra weight in the analysis.
5. Apply an outlier test to the triplets of emissions tests, delete any test found to be an outlier, and then use the arithmetic mean of the remaining tests at each testing point. This procedure deals with potential outliers among the emissions tests, and removes the problem of vehicles with third tests getting extra weight in the analysis. However, if there are no outliers among the three tests, and they have the same underlying test-to-test (error) variance, then the means from three tests will have smaller standard errors, and the statistical model will not take this into account.

An additional difficulty with using the individual tests is that SAS software cannot easily be used to develop mixed models with repeated measure effects (i.e., within vehicle correlations) when there are different test structures for each subject vehicle.

After some consideration it was decided to apply option 5, as used in the Auto/Oil Program. There were eight triplets (per pollutant) in the database supplied by EPA. For each triplet, the maximum deviation from the sample mean is divided by the pooled standard deviation estimate (computed from all triplets and pairs of emissions tests) and compared to the tabulated critical value of 1.9400, which is based on 90 degrees of freedom for the other triplets and all pairs, and

using a 5 % test. The 5 % level refers to the probability (for each triplet) of finding a triplet to be an outlier when all three trials are “valid.” No adjustment was made for the fact that eight comparisons were made, so that the probability of wrongly detecting at least one outlier test is greater than 5 % (a multiple comparisons problem). Note that the Auto/Oil Program used a 10 % level, but their protocol excluded a trial only if the deviations from the mean divided by the pooled standard deviation exceeded the critical value for two or three of the pollutants HC, CO, or NO<sub>x</sub>.

The following outlier trials were deleted:

NO<sub>x</sub>: Accord, vehicle 2, phase 0, trial 1 (significant at 5 %, but not 1 %)

HC: F150, vehicle 1, phase 1, trial 2 (significant at 1 %)

CO: Accord, vehicle 4, phase 1, trial 1 (significant at 5 %, but not 1 %)

The phase means were computed after deletion of these outliers, and the natural logarithms of the phase means were used as the dependent variable (experience with vehicle emissions data supports the assumption of approximate log-normality for vehicle emissions).

### **Phase Mean Outliers**

This section describes the outlier screening methods that were used in the model development below. Two approaches were used. One approach was based on studentized deletion residuals from the GLM models. A second approach was based on the Rosner (1975) RST multiple outlier test applied to the studentized deletion residuals for the GLM models and was applied to the (unstudentized) residuals for the MIXED models. The Rosner (1975) test is based on the approximation that these residuals are independent and have identical normal distributions.

The initial model development was based on a general linear model, GLM, i.e., with all effects treated as fixed effects. In SAS, the GLM procedure allows studentized deletion residuals to be output. These are the residuals (observed value minus predicted value) divided by their estimated standard deviations, where the standard deviations are estimated using all phase means except for the observation in question. Studentized deletion residuals greater than 2.5 in absolute value were treated as potential outliers. This calculation is not currently available for the MIXED procedure, used to fit models with fixed and random effects.

The adapted Rosner (1975) RST test was employed for both the GLM and MIXED models. The advantages of this approach are that the method allows for the possibility of more than one extreme residual and also provides significance levels for the outlier tests. However, the method is approximate because the residuals are not independent and because a large sample approximation was used to derive the critical values. In addition, the residuals from the MIXED model are not identically distributed.

The method was based on the RST algorithm developed by Rosner (1975), as described by Barnett and Lewis (1994; pages 235–6). To avoid complex table look-up procedures, the method (developed for a previous project) used a large sample approximation. Probably the most crucial issue is that typical rules that reject logged values two or three standard deviations away from the mean ignore the fact that the log-normal assumption implies that a certain percentage of values would be expected to lie outside those limits (roughly 5 percent and 0.3 percent), so those naive approaches tend to lead to too many rejected values.

The modified RST algorithm is described in the next paragraph. The algorithm is complex to describe, but not too difficult to apply. The algorithm uses as input a set of datapoints assumed to come from a normal distribution apart from up to  $K$  "contaminated" extreme values (too high or too low).  $K$  is estimated as part of the algorithm. For the residuals from the MIXED models, the underlying unknown mean and standard deviation were estimated robustly by winsorization: Each of the  $K$  highest values are temporarily replaced by the  $(K+1)$ th highest value, and each of the  $K$  lowest values are replaced by the  $(K+1)$ th lowest value, and then the sample mean and standard deviation is calculated from the adjusted data. The winsorization method is designed to robustly estimate the mean and variance in the presence of potential outliers. For the studentized deletion residuals from the fitted general linear models (GLM procedure), we made the approximation that the underlying mean and variance of the studentized residuals are 0 and 1, respectively.

The multiple outlier detection algorithm assumes up to  $K$  possible extreme outlier values for the sample of  $N$  values. In the first step, the value of  $K$  is estimated using a maximum gap approach. First, the set of unsigned residuals is arranged in increasing order. Starting from the lowest unsigned residual, the differences between consecutive unsigned residuals are compared and the highest such gap is found among the 10 percent of the data furthest from zero; all values above that gap are presumed to be potential outliers.  $K$  is estimated to be the number of potential outliers, but the value of  $K$  is increased if necessary to include multiple instances of the same large unsigned residual.

In the second step, the mean value of zero used for the maximum gap algorithm is replaced by a more robust estimate, and the  $K$  values furthest from the underlying mean are standardized to give the number of standard deviations away from the mean,  $T = |x - \text{MEAN}|/\text{SD}$ . The values of MEAN and SD are either the winsorized mean and standard deviation, for the residual analyses, or 0 and 1, for the studentized deletion residual analyses. Let  $T_1$  be the highest value of  $T$ , which corresponds to the observation furthest from MEAN. Let  $T_2$  be the second highest value of  $T$ . Similarly,  $T_j$ , for  $j \leq K$  is the  $j$ th highest value of  $T$ .

The third step defines the outlier detection rule. Let  $X_j$  be critical values to be defined shortly, for  $j = 1$  to  $K$ . If  $T_K > X_K$ , then the observations corresponding to  $T_1, T_2, \dots, T_K$  are all assumed to be outliers, which completes the algorithm. Otherwise, if  $T_{K-1} > X_{K-1}$ , then the observations corresponding to  $T_1, T_2, \dots, T_{K-1}$  are all assumed to be outliers, which ends the algorithm. The values of  $T_j$  are sequentially compared with the critical values in decreasing order of  $j$ . For the

final comparison, if  $T_1 > X_1$ , then only the observation corresponding to  $T_1$  is presumed to be an outlier. Otherwise, no outliers are detected.

Let  $\alpha$  be the required overall significance level (e.g. 5 percent, or 10 percent). The critical values  $X_j$  are defined as the solutions of the equations  $\text{Prob}(T_j > X_j) = \alpha/K$ , calculated under the null assumption of no outliers. By the Bonferroni inequality, it follows that the probability that one or more values of  $T_j$  exceed their critical value is at most  $\alpha$ , which is also an upper bound to the probability of detecting one or more outliers when none are present. If there are no outliers, and if the sample sizes are large enough that MEAN and SD can be assumed to equal the true mean and standard deviation, then  $T_j$  has the same distribution as the  $j$ th highest, regardless of sign, from  $N$  values independently generated from a standard normal distribution. It follows that  $\text{Prob}(T_j > X_j)$  equals the binomial probability,  $B(j, X_j)$ , of  $j$  or more successes in  $N$  independent trials, where the probability of success equals  $P(X_j)$ .  $P(x)$  is defined as the probability that a single unsigned standard normal variable exceeds  $x$ . We obtain the following formulae:

$$P(x) = \int_{|z|>x} \frac{1}{\sqrt{2\pi}} e^{-0.5z^2} dz,$$

$$P(T_j > x) = B(j, x) = \sum_{r=j}^N \binom{N}{r} [P(x)]^r [1 - P(x)]^{N-r},$$

$$B(j, X_j) = \alpha / K.$$

Rather than solving the last equation for the critical value, it is equivalent, and more convenient to use the result that  $T_j$  exceeds the critical value if and only if  $B(j, T_j)$  is less than or equal to  $\alpha/K$ . For the following analysis an experiment-wise significance level of 5 % was used for the RST test.

Under both these approaches, outliers are defined for specific combinations of vehtype, vehicle and phase, so that if a vehicle has an outlier for a single phase, the other phase means are still used. This is reasonable since it is plausible that the problem causing the outlier was a short term problem for that set of FTP tests. Nevertheless, after developing the final models, we examined the sensitivity of the models to alternative outlier protocols including a protocol where all data for a vehicle with at least one phase mean outlier are deleted.

### Intercept and Slope Terms - General Linear Model

The initial model development was based on a general linear model approach, i.e., with all terms treated as fixed effects. Because of the large variability between vehicles, even of the same vehicle type, this model included a vehicle-specific intercept; equivalently, the model had terms for the overall intercept, vehtype and the vehicle\*vehtype interaction. Vehtype and vehicle are

both classification variables. In addition, the basic model was forced to include direct and carryover fuel effects, as defined above, with separate values for each test fuel. Clearly some fuel effects terms need to be included in the model in order to estimate fuel effects on emissions and their statistical significance. Furthermore, it is important to include fuel effects in the initial model because the statistical significance of other main and interaction effects will depend upon which other terms are included in the model. Subsequent analyses determine the need for including interactions between these fuel effects and other effects or for combining the alternating and continuous fuel effects.

Starting with the initial model we then considered a sequence of more refined models including adjustments for the miles effects, i.e., the emissions deterioration on the reference fuel. Adding the miles effect gave a new model with a p-value (significance level) of 0.0191 for miles, so this term was included. Adding a further term for the interaction between miles and vehtype (different slopes for different vehicle types) gave a new model with p-values of 0.0084 for miles and 0.0060 for miles\*vehtype, so this interaction was included. Finally, adding a further interaction between miles and vehicle gave p-values of 0.0010, 0.0004, and 0.0087 for miles, miles\*vehtype, and miles\*vehicle, respectively. Thus the interaction miles\*vehicle should also be included. All these p-values are incremental Type I p-values for adding that term in the presence of the previously added terms.

The above analysis was carried out using all 84 phase means. However, outlier phase means were identified for each of the statistical models, as listed here: The base model had the studentized deletion residual (SDR) outlier F150-4-0, using the abbreviation vehtype-vehicle-phase, but there were no RST outliers at the 5 % level. The miles model had SDR outliers F150-4-0 and F150-4-2 and a single RST outlier F150-4-0. The miles and miles\*vehtype model had SDR outliers F150-4-0, LeSabre-1-0, and F150-2-0 and RST outliers F150-4-0 and LeSabre-1-0. The miles miles\*vehtype, miles\*vehicle model had SDR outliers Accord-2-1, Accord-2-0, Accord-2-2, F150-4-2, F150-4-0, and F150-4-1, which were also the RST outliers. The F150-4-0 phase mean is an outlier for most of these analyses and should be considered for exclusion. One approach would be to delete that phase mean and all the other phase mean outliers identified in these analyses. However, removing too many outliers makes the analysis much less representative of the available data. Furthermore, if the most extreme outlier is removed and the model is refitted, it is often the case that the remaining outliers are no longer outliers for the refitted model. For these reasons, only the F150-4-0 outlier was deleted and the model was refit to the remaining 83 phase means.

Using the 83 phase means, the base model had a SDR outlier for LeSabre-3-0 and no RST outliers. The miles model had a p-value of 0.0042 for miles and no outliers. The miles miles\*vehtype model had p-values of 0.0018 and 0.0114, and LeSabre-1-0 was a SDR and RST outlier. The miles miles\*vehtype miles\*vehicle model had p-values of 0.0002, 0.0012, and 0.0162, respectively, and the SDR and RST outliers were for Accord-2-0, Accord-2-1, and Accord 2-2. This sequence of fitted models again suggests that all three slope terms should be included in the model, but also suggests that some of the phase means were outliers. Although

one can argue for eliminating either the LeSabre-3-0, the LeSabre-1-0, or all three phase means for Accord-2, we found later that the LeSabre-1-0 phase mean was the only outlier when the corresponding initial mixed model was fitted to the same set of 83 phase means (see below). Therefore, the LeSabre-1-0 outlier was deleted and the model was refit to the remaining 82 phase means.

Using the 82 phase means, the base model had a SDR outlier for LeSabre-3-0 and no RST outliers. The miles model had a p-value of 0.0007 for miles and no outliers. The miles\*vehtype model had p-values of < 0.0001 and 0.0004, a SDR outlier for Accord-3-2, and no RST outliers. The miles miles\*vehtype miles\*vehicle model had p-values of < 0.0001, 0.0002, and 0.0788, respectively, and the SDR and RST outliers were for Accord-2-0, Accord-2-1, and Accord 2-2. This sequence of fitted models suggests that the miles and miles\*vehtype terms should be included in the model, but the miles\*vehicle interaction term may not be sufficiently statistically significant for inclusion. The sequence of models also suggests the possibility of additional outliers. However, at this point it was decided that further outlier deletions would be counterproductive, since the resulting models would be even less representative of the entire database.

### **Intercept and Slope Terms - Mixed Model**

At this point in the model development, the initial general linear model was redefined as a mixed model, with fixed and random effects. A fixed effect is a constant model parameter. A random effect has unknown values for each subject that are assumed to have been generated from a certain distribution, typically assumed to be normal, with an underlying mean of zero. In this case the subjects are individual vehicles. Effects can be assumed random if the subjects can be thought of as a random sample from a population. In this situation, the vehtypes cannot be thought of as a random sample from the national fleet, since they were deliberately selected to cover a range of emissions certification levels; the national proportions of vehicles with those levels are not equal. (The distributions are also unlikely to be normal, although the normality is not essential to the issue of fixed or random effects; normality assumptions are only used for the model fitting and for estimating significance levels). Thus the vehtype effect is a fixed effect. However, the four vehicles selected for each vehtype can reasonably be assumed to be a random sample from that vehtype. Thus the vehicle (nested within the vehtype) can be assumed to be a random effect.

All the mixed models in the following analyses were fitted using the maximum likelihood approach. This approach assumes that the random effects (including the residuals) are normally distributed. Alternative approaches using sums of squares do not have the attractive asymptotic (large sample) properties of maximum likelihood estimates and can lead to impossible negative variance estimates. The restricted maximum likelihood method has similar asymptotic properties and tends to reduce the bias of the estimated covariance parameters. However, restricted maximum likelihood has the disadvantage that restricted log-likelihood or Akaike Information Criterion (AIC) comparisons between models with different sets of fixed effects terms are

invalid. Restricted maximum likelihood and maximum likelihood estimates are usually very similar.

Based on the results of the last section, the initial mixed model had fixed effects for vehtype (seven vehtype intercepts, or, equivalently, an overall intercept and six independent vehtype adjustments), miles (the overall slope), miles\*vehtype (slopes for each vehtype), and the four fuel effects (direct and carryover effects for each test fuel). The initial mixed model also had random effects for the intercept and miles, with the vehicles as the subjects (more specifically, the subject is vehicle(vehtype), i.e., vehicle nested within vehtype). These random effects are the random variation of the intercepts and slopes across the four vehicles for each vehtype, since the mean values for each vehtype are fixed effects. In this model the variances of the intercepts and slopes are the same for all vehtypes. The miles random effect may not be required since the equivalent GLM model with the two outliers removed showed that this effect had a p-value of only 0.08.

Fitting this mixed model to all 84 phase means showed that the F150-4-0 and LeSabre-1-0 phase means were both RST outliers (SDR outliers are not available for the MIXED procedure). Fitting this mixed model to the 83 phase means excluding F150-4-0 showed that the LeSabre-1-0 phase mean was a RST outlier. Fitting this mixed model to the 82 phase means excluding F150-4-0 and LeSabre-1-0 found an additional eight RST outliers: Accord-2-1, F150-2-0, Accord-3-2, Accord-1-0, Accord-2-2, F150-2-1, Escort-4-0, and Caravan-2-2. However, to obtain a representative model of the bulk of the data, it was decided to retain all but the F150-4-0 and LeSabre-1-0 phase means for the rest of the model development.

The initial model (for the 82 phase means) had a  $-2 \times \log$ -likelihood value of -113.5 with a total of 21 fixed and random effect parameters. To evaluate the need to include the miles random effect, which was not significant at the 5 % level for the corresponding GLM model, the same model without the miles random effect was fitted. The alternative, simpler model had a  $-2 \times \log$ -likelihood value of -113.2 with a total of 20 parameters. The difference between the  $-2 \times \log$ -likelihood values is only 0.3, with 1 degree of freedom, so the chi-square likelihood ratio test shows that the miles random effect in the more complicated model is not statistically significant at the 5 % level. Thus the miles random effect was dropped from the model. For later reference we shall refer to this model without the miles random effect as model A. The estimated fuel effects for model A are shown in Table 3, below.

## **Fuel Effects**

This section presents and discusses the analysis of possible interactions between the fuel effects and miles, vehtype, or vehicle. Also discussed is the validity of using data from phase 2 in relation to aliasing of effects and the representativeness of the fuel effects for the national fleet. Determinations of possible model simplifications by eliminating carryover effects or combining the two test fuels are made later in the model development.

## **Phase 2 Data**

As discussed in the section “Summary of Statistical Models,” the fuel effects are the direct fuel effect and the carryover effect.

The direct fuel effect applies at 8,000 and 16,000 miles and estimates the change in emissions attributable to the test fuel over the previous 8,000 miles, compared to the expected change had reference fuel been used. The direct fuel effect at 8,000 miles applies to vehicles 1 and 3 only. The direct fuel effect at 16,000 miles applies to vehicles 2 and 4 only. It is possible that the direct effect varies by phase, but such an interaction effect is hard to interpret. The difference between phase 1 and phase 2 is the extra 8,000 miles on the reference fuel, but this is unrelated to differences in the total mileage accumulation (from zero miles): Table 2 shows that for each vehtype, the total mileage accumulation for vehicles 2 and 4 at the beginning of phase 2 can be equal, less, or more than the total mileage accumulation for vehicles 1 and 3 at the beginning of phase 1. A phase\*direct interaction might be attributable to differences between the reference fuel and the fuels used prior to the study. It also might be attributable to the effect of the driving cycle used for the study mileage accumulation. In either case, one would not be able to use the study results to estimate impacts on the national fleet because the phase effect is associated with the experimental study and cannot be defined for an in-use vehicle. The possible interaction between the direct fuel effect and phase is evaluated below.

If the direct effect depends on the total mileage accumulation, rather than the phase, then the results of the study cannot easily be extrapolated to the national fleet, because the fleet has a wide variety of mileage accumulations by vehtype and model year. It would be necessary to model and estimate the direct effect as a function of the mileage accumulation.

The “statistical” carryover for each test fuel is defined as the expected difference between the emissions at the end of phase 2 for the sequence Test-R (i.e., test fuel followed by reference fuel) compared to the sequence R-Test. Since the expected intercept depends only on vehtype, and not on the vehicle (the vehicle intercept is a random effect), this is the expected difference, at the end of phase 2, between vehicles 1 and 2 for the continuous test fuel or between vehicles 3 and 4 for the alternating test fuel, for vehicles in the same vehtype.

We assume that if reference fuel had been used on the same vehicle for 16,000 miles, the expected log(NO<sub>x</sub>) emissions increment over phase 1 would be the same as that over phase 2. This assumption would not be true if the log(NO<sub>x</sub>) on reference fuel was a non-linear function of mileage accumulation, and in that case the phase 1 and 2 increments would both depend on the initial mileage accumulation at the beginning of the study. Table 2 above shows that the initial mileage accumulations for all the test fleet vehicles vary widely within each vehtype, by up to 41K miles. Without the assumption, a more complicated model formulation with non-linear mileage effects, possibly varying by vehtype, could adjust for these effects. Such a model could account for the initial mileage accumulation impacts on the reference and test fuel increments, avoiding the aliasing of fuel effects with the initial mileage effect. However, the assumption should hold approximately, even if the log(NO<sub>x</sub>) is non-linear, since the emissions slope for a given vehicle at a given initial mileage will not vary very much over 16,000 miles. Under the

assumption, the “statistical” carryover effect for the reference fuel in phase 2 is not aliased with a miles or phase effect, and can be estimated from the study data. (Aliasing would mean that the two effects cannot be distinguished using the study data). The “engineering” carryover effects can be estimated as the sum of the phase 2 direct effects and the “statistical” carryover effects.

In summary, the above argument shows that it is valid to use the phase 2 data in the analysis even though a variant of the two stage crossover experimental design has been used and carryover cannot be assumed negligible. The phase 2 data for the reference fuel provides estimates of the carryover, when compared with the phase 1 data. The phase 2 data for the test fuels provides estimates of the direct effects in phase 2. These will be the same as the direct effects in phase 1 if there is no interaction between direct effects and miles (from start of study).

### Interactions With Vehtype

Starting with the mixed model A developed in the last main section (the initial mixed model without the miles random effect), additional fuel effect interactions with vehtype were added to the model. Adding the fixed effects direct\*vehtype and applying the approximate statistical F tests for fixed effects showed this set of effects had a p-value of 0.61 with 12 degrees of freedom (6 for each test fuel). Adding both the direct\*vehtype and carryover\*vehtype interaction effects gave p-values of 0.22 and 0.02, each with 12 degrees of freedom, respectively. Adding only the carryover\*vehtype interaction effects gave a p-value of 0.04. These results suggest that the carryover effect varies with vehtype, although the direct effect is constant across vehtypes.

The improvements in the model fit from adding these fuel\*vehtype interactions come at the cost of significantly increasing the number of model parameters. The direct\*vehtype and carryover\*vehtype interactions each add 12 model parameters, giving a total of 32 parameters after adding one set of interactions and a total of 44 parameters after adding both sets of interactions. There are only 82 datapoints (after excluding the two outliers). The Akaike Information Criterion (AIC) is a measure of fit that adjusts for the possibility of over-fitting by penalizing the likelihood for the number of fitted parameters. The AIC is an estimate of the “entropy information” in the data, and is a measure of the difference between the fitted model and the unknown true model. AIC is defined by  $-2 \times \log\text{-likelihood} + 2 \times \text{number of free parameters}$ , so that lower AIC values suggest better fitting models. (Note that some texts define the AIC without the factor of 2 and change the sign, so that higher AIC values suggest better fitting models). For model A,  $\text{AIC} = -73.2$ . Using this definition, the AIC values increased from the original model when any of the interaction terms were added, which is evidence of overfitting.

Another reason not to include fuel\*vehtype interactions in the model is the representativeness of the results. If the fuel effects depend upon the vehtype, then they cannot simply be extrapolated to the national fleet, since the vehtypes were not a random sample of vehtypes. If the fuel effects do not depend on the vehtype then they can be extrapolated to the national fleet.

For these reasons it was decided not to add the fuel\*vehtype interactions to the model, but to reconsider this issue later in the development if a model with fewer parameters could be fitted to the data (for example by combining the effects of the two test fuels).

### Interactions with Miles

Since carryover only applies in phase 2, the carryover effects cannot have interactions with miles. Adding to model A the direct\*miles interaction terms (one term for each test fuel) gave a p-value of 0.0579, not statistically significant at the 5 % level, although close to being significant. Another reason to not include these terms is that this interaction is, by definition, the interaction with the phase, rather than with the total accumulated mileage. As discussed above, this interaction is hard to interpret, and such an interaction would make it impossible to use the study results to estimate impacts on the national fleet.

### Interactions with Vehicle

If the fixed fuel effects are assumed to be independent of the vehtype, it is still plausible that the fuel effects vary randomly between vehicles, around a mean value of zero. To test this assumption, three variants of model A were fitted with additional random effects: either 1) the direct effects only, 2) the carryover effects only, or 3) all four fuel effects. The decreases in the  $-2 \times \log$ -likelihood values were 0.0, 1.8, and 1.8. Formal likelihood ratio tests could not be carried out because the alternative models were on the boundary of the parameter space, having zero estimated variances for some or all of these random effect terms. However the AIC showed that these models did not improve upon model A.

### Representativeness

In this section the various analyses have led to the retention of model A, because possible interactions were either not statistically significant, led to an overfitted model (based on AIC), and/or by assumption, because they led to a model that could not be used to represent the national fleet. Since the fuel effects in model A are independent of vehtype, vehicle, and miles, this model can be applied to estimate emissions impacts on the national fleet.

### **Variance Homogeneity**

A further model enhancement to model A was investigated to evaluate the possibility that the random effect variances depend upon the vehtype. In this model the variance of the intercept was assumed to vary with vehtype instead of being a constant. The revised model had a  $-2 \times \log$ -likelihood value of -120.5 (a reduction of 7.3) with a total of 26 degrees of freedom, so the improvement was not statistically significant according to the likelihood ratio test. Model A was retained.

### **Repeated Measures**

Additional enhancements to model A consider possible correlations between measurements on the same vehicle (over the three phases). Note that the intercept random effect is a vehicle-specific random variable that is constant for the three phases on that vehicle and therefore introduces a constant correlation between phases. These alternative repeated measures models add further complexity to the correlation structure.

Various models were evaluated. The first model was a variance components model grouped by vehtype, which, in effect, assumes that the residual variance depends on the vehtype. This model, B, was almost as good as model A since the  $-2 \times \log$ -likelihood value was reduced to -124.5, a reduction of 11.3 with 6 degrees of freedom (a total of 26 model parameters); this reduction is statistically significant at the 10 % level but not at the 5 % level. The AIC was only slightly higher, i.e., only slightly worse (-72.6 instead of -73.2). The other models were as follows:

- An autoregressive model such that the correlation between phases is  $\rho^{|\text{difference in miles}|}$ . For this model the  $-2 \times \log$ -likelihood value was reduced to -113.9, a reduction of 0.7 with 1 degree of freedom; this reduction is not statistically significant at the 10 % level. The AIC was -72.0 (higher). Both results suggest model A is preferred.
- The same autoregressive model with the additional feature that the variances depend upon the phase. For this model the  $-2 \times \log$ -likelihood value was reduced to -116.9, a reduction of 3.7 with 3 degrees of freedom; this reduction is not statistically significant at the 10 % level. The AIC was -71.0 (higher). Both results suggest model A is preferred.
- An unstructured model with arbitrary phase variances and arbitrary covariances between measurements in different phases. For this model the  $-2 \times \log$ -likelihood value was reduced to -116.9, a reduction of 3.7 with 6 degrees of freedom; this reduction is not statistically significant at the 10 % level. The AIC was -65.0 (higher). Both results suggest model A is preferred.
- An unstructured model with arbitrary phase variances but zero covariances between measurements in different phases. (Note that the full model has constant covariances between phases because of the random vehicle intercept. This description is only for the additional repeated measures structure). For this model the  $-2 \times \log$ -likelihood value was reduced to -116.7, a reduction of 3.5 with 2 degrees of freedom; this reduction is not statistically significant at the 10 % level. The AIC was -72.6 (higher). Both results suggest model A is preferred.

From these results it was decided to retain models A and B for further model development. The estimated fuel effects for models A and B are shown in Table 3 below.

## Fuel Equivalence

To examine the possibility that the test fuel effects were not statistically significantly different, simplified versions of models A and B were fitted with the two test fuels assumed to have equal effects.

Compared to model A, the version of model A with equal test fuel effects had a  $-2 \times \log$ -likelihood value of -111.8, a difference of 1.4 with 2 degrees of freedom; this difference is not statistically significant at the 10 % level. The AIC was -75.8 (lower). Furthermore, by choosing a suitable linear combination of the four fuel effects that equals zero if the two test fuels have equal effects, the significance level of this contrast provides an alternative direct test of the fuel equivalence. The p-value was 0.4926. All these results show that the version of model A with equal fuel effects is an improvement over model A. This model is referred to as model 1 below.

Compared to model B, the version of model B with equal test fuel effects had a  $-2 \times \log$ -likelihood value of -123.9, a difference of 0.6 with 2 degrees of freedom; this difference is not statistically significant at the 10 % level. The AIC was -75.8 (lower). The contrast p-value was 0.6858. All these results show that the version of model B with equal fuel effects is an improvement over model B. This model is referred to as model 2 below.

Comparing models 1 and 2, the AIC values are identical and the  $-2 \times \log$ -likelihood difference is 11.1 with 6 degrees of freedom, which is significant at the 10 % level but not at the 5 % level.

On the basis of these results, the final models were selected to be models 1 and 2. The fuel effects for models 1 and 2 are shown in Table 3. Table 3 also includes the individual fuel effects estimated in models A and B. The fuel effects are expressed as direct effects, “engineering” carryover effects, and “statistical” carryover effects. Also included in these tables are standard errors for the fuel effects, p-values, and asterisks to indicate significant effects at the five percent level. These results show that in most cases the direct fuel effects are small and not statistically significant, but the “engineering” and “statistical” carryover fuel effects are larger and typically statistically significant.

Even though the results for models A and B show that the direct effects for the continuous test fuel were stronger than for the alternating fuel (agreeing with intuition), these differences were not statistically significant. Since the carryover effects for the two test fuels were stronger, statistically significant, and similar, the overall fuel equivalence analysis showed that the two fuels could be treated as having equal effects.

<b>Table 3. Final models for estimating NOx emissions impacts (changes in the natural logarithm of emissions) attributable to the fuel additive.</b>							
Model	Fuels	Direct effect	Standard error of direct effect	“Statistical” carryover effect	Standard error of “statistical” carryover effect	“Engineering” carryover effect	Standard error of “engineering” carryover effect
A	T	-0.045 (0.32)	0.045	<u>-0.111</u> (0.02)	0.047	<u>-0.156</u> (0.03)	0.070
A	A	-0.001 (0.99)	0.046	<u>-0.114</u> (0.02)	0.046	-0.115 (0.12)	0.072
1	T, A	-0.022 (0.59)	0.042	<u>-0.112</u> ( $< 0.01$ )	0.037	<u>-0.135</u> (0.05)	0.066
B	T	-0.043 (0.26)	0.037	-0.063 (0.11)	0.038	-0.106 (0.08)	0.058
B	A	-0.016 (0.67)	0.037	<u>-0.089</u> (0.02)	0.038	-0.105 (0.08)	0.059
2	T, A	-0.025 (0.46)	0.034	<u>-0.072</u> (0.02)	0.030	-0.097 (0.08)	0.054

Notes:

- (1) Models A and 1 have equal error variances for all vehicle types. Models B and 2 have different error variances for each vehicle type.
- (2) The FTP mean outliers for phase 0, vehtype F150, vehicle 4 and phase 0, vehtype LeSabre, vehicle 1 have been deleted.
- (3) Statistically significant effects at the five percent level are underlined. P-values (significance levels) are shown in parentheses.

### SAS Code

To clarify the formulations used for models 1 and 2, the following SAS code can be used to fit these models to the phase mean dataset “data”, after removing the two outlier phase means. The value of log(NOx) was assigned the variable name logemis. The direct and carryover variables are defined as:

direct =        1 for phase 1, vehicles 2 and 4  
                   1 for phase 2, vehicles 1 and 3  
                   0 otherwise

carryover =    1 for phase 2, vehicles 2 and 4

0 otherwise

### SAS Code for Model 1

```
proc mixed data=data method=ml ic maxiter=50;
class vehtype vehicle phase direct carryover;
model logemis= vehtype miles miles*vehtype direct carryover / solution cl;
random intercept / subject=vehicle(vehtype);
run;
```

### SAS Code for Model 2

```
proc mixed data=data method=ml ic maxiter=50;
class vehtype vehicle phase direct carryover;
model logemis= vehtype miles miles*vehtype direct carryover / solution cl;
random intercept / subject=vehicle(vehtype);
repeated phase / subject=vehicle(vehtype) type = vc group=vehtype;
run;
```

### **Further Analysis of Fuel Interactions**

As a further check of the final models 1 and 2, possible interactions between the fuel effects and vehtype or miles were added to these models (which have fewer parameters than model A).

For model 1, adding the interactions between fuel and vehtype decreased the  $-2 \times \log$ -likelihood value to -126.5, a reduction of 14.7 with 12 degrees of freedom; this reduction is not statistically significant at the 10 % level. The AIC was -66.5 (higher). Also for model 1, adding the interaction between fuel and miles decreased the  $-2 \times \log$ -likelihood value to -112.5, a reduction of 0.7 with 1 degree of freedom; this reduction is not statistically significant at the 10 % level. The AIC was -74.5 (higher).

For model 2, adding the interactions between fuel and vehtype decreased the  $-2 \times \log$ -likelihood value to -143.9, a reduction of 20.0 with 12 degrees of freedom; this reduction is not statistically significant at the 5 % level but is statistically significant at the 10 % level. The AIC was -71.9 (higher). Also for model 2, adding the interaction between fuel and miles decreased the  $-2 \times \log$ -likelihood value to -125.7, a reduction of 1.8 with 1 degree of freedom; this reduction is not statistically significant at the 10 % level. The AIC was -75.7 (higher).

These analyses show that adding fuel\*vehtype or fuel\*miles interactions to the final models did not significantly improve the fit. Another way of looking at these results is that the analysis based on the final models shows that the fuel effects can be assumed not to depend upon the vehtype or the phase. As discussed above, this provides support to the contention that the model results are representative of effects in the national fleet.

## **FINAL MODELS FOR NO<sub>x</sub>, HC, and CO**

The structure of models 1 and 2 was developed using an analysis of the NO<sub>x</sub> data only. For comparison, the same models were fitted to the HC and CO phase means (after deleting the respective triplet outliers, as described above). The results are shown in Table 4. Table 4 also compares results for different outlier approaches.

The final models 1 and 2 use all of the data except for the phase means F150-4-0 and LeSabre-1-0. As for NO<sub>x</sub>, the final models (1 and 2) for HC and CO show strong negative carryover effects (except for “engineering” carryover for CO in model 2) and show small direct effects (except for HC model 2).

The alternative models 1A-1F and 2A-2F are variants of models 1 and 2, respectively, with different approaches for dealing with the outliers. Alternative options are to exclude none, one, or both phase means, or to exclude one or both of the vehicles that had outlier phase means (F150-4 or LeSabre-1). For HC and CO, the results were not very sensitive to any of the outlier protocols. This might be expected since the model outliers were chosen based on fitting models to the NO<sub>x</sub> data and therefore may not be outlier values for HC or CO.

Models 1A and 2A use all of the data. As might be expected, for NO<sub>x</sub>, the estimated fuel effects for these models are the most different from models 1 and 2. This confirms their exclusion as statistical outliers. Models 1B and 2B do not use any data from the outlier vehicles (i.e., the vehicles with a phase mean outlier). The NO<sub>x</sub> results for these two models are relatively similar to the model 1 and 2 results, which suggests that the other phase means for those vehicles are not outliers from the statistical models. Models 1C and 2C do not use the outlier phase mean from the F150-4, whereas models 1D and 2D exclude that vehicle’s data entirely. Similarly, models 1E and 2E do not use the outlier phase mean from the LeSabre-1, whereas models 1F and 2F exclude that vehicle’s data entirely. For the C, D, E, and F models, the estimated fuel effects on NO<sub>x</sub> differ from the final models (1 and 2), but not by as much as the models 1A and 2A. Again this is as expected since each of those 8 models was based on data including one outlier phase mean.

**Table 4. Final models for estimating emissions impacts (changes in the natural logarithm of emissions) attributable to the fuel additive. Proposed models in bold.**

Pollutant	Model	Direct effect	Standard error of direct effect	“Statistical” carryover effect	Standard error of “statistical” carryover effect	“Engineering” carryover effect	Standard error of “engineering” carryover effect
NO <sub>x</sub>	<b>1</b>	<b>-0.022</b>	<b>0.042</b>	<b><u>-0.112</u></b>	<b>0.037</b>	<b><u>-0.135</u></b>	<b>0.066</b>
	1A	0.038	0.052	-0.060	0.047	-0.022	0.081
	1B	-0.013	0.043	<u>-0.116</u>	0.039	-0.130	0.067
	1C	0.012	0.047	-0.078	0.042	-0.066	0.074
	1D	0.008	0.047	<u>-0.089</u>	0.043	-0.081	0.074
	1E	0.013	0.048	-0.087	0.043	-0.074	0.075
	1F	0.030	0.049	-0.076	0.044	-0.046	0.076
	<b>2</b>	<b>-0.025</b>	<b>0.034</b>	<b><u>-0.072</u></b>	<b>0.030</b>	<b>-0.097</b>	<b>0.054</b>
	2A	-0.005	0.034	-0.042	0.031	-0.047	0.055
	2B	-0.019	0.034	<u>-0.067</u>	0.031	-0.086	0.055
	2C	-0.014	0.034	-0.049	0.031	-0.063	0.054
	2D	-0.018	0.034	-0.054	0.031	-0.072	0.055
	2E	-0.016	0.034	<u>-0.064</u>	0.031	-0.080	0.055
	2F	-0.005	0.034	-0.053	0.031	-0.058	0.055
HC	<b>1</b>	<b>0.045</b>	<b>0.027</b>	<b><u>0.091</u></b>	<b>0.025</b>	<b><u>0.136</u></b>	<b>0.044</b>
	1A	0.030	0.028	<u>0.077</u>	0.026	<u>0.108</u>	0.045
	1B	<u>0.057</u>	0.027	<u>0.096</u>	0.025	<u>0.153</u>	0.043
	1C	0.033	0.029	<u>0.079</u>	0.026	<u>0.112</u>	0.045
	1D	0.036	0.029	<u>0.075</u>	0.026	<u>0.111</u>	0.046
	1E	0.042	0.027	<u>0.089</u>	0.025	<u>0.130</u>	0.044
	1F	0.051	0.027	<u>0.098</u>	0.024	<u>0.148</u>	0.042

**Table 4. Final models for estimating emissions impacts (changes in the natural logarithm of emissions) attributable to the fuel additive. Proposed models in bold.**

Pollutant	Model	Direct effect	Standard error of direct effect	“Statistical” carryover effect	Standard error of “statistical” carryover effect	“Engineering” carryover effect	Standard error of “engineering” carryover effect
HC	<b>2</b>	<b><u>0.063</u></b>	<b>0.022</b>	<b><u>0.091</u></b>	<b>0.020</b>	<b><u>0.155</u></b>	<b>0.036</b>
	2A	<u>0.059</u>	0.022	<u>0.088</u>	0.020	<u>0.148</u>	0.036
	2B	<u>0.069</u>	0.022	<u>0.087</u>	0.020	<u>0.155</u>	0.035
	2C	<u>0.062</u>	0.023	<u>0.090</u>	0.020	<u>0.152</u>	0.036
	2D	<u>0.066</u>	0.022	<u>0.085</u>	0.020	<u>0.151</u>	0.036
	2E	<u>0.061</u>	0.022	<u>0.089</u>	0.020	<u>0.150</u>	0.035
	2F	<u>0.062</u>	0.022	<u>0.090</u>	0.020	<u>0.151</u>	0.035
CO	<b>1</b>	<b>0.043</b>	<b>0.041</b>	<b><u>0.127</u></b>	<b>0.037</b>	<b><u>0.167</u></b>	<b>0.065</b>
	1A	0.023	0.043	<u>0.107</u>	0.039	0.129	0.068
	1B	0.059	0.042	<u>0.139</u>	0.038	<u>0.198</u>	0.066
	1C	0.025	0.044	<u>0.108</u>	0.039	0.133	0.068
	1D	0.025	0.044	<u>0.108</u>	0.040	0.133	0.070
	1E	0.040	0.040	<u>0.125</u>	0.037	<u>0.165</u>	0.065
	1F	0.055	0.041	<u>0.137</u>	0.036	<u>0.192</u>	0.064
	<b>2</b>	<b>-0.008</b>	<b>0.028</b>	<b><u>0.096</u></b>	<b>0.025</b>	<b>0.088</b>	<b>0.044</b>
	2A	-0.014	0.027	<u>0.093</u>	0.024	0.080	0.043
	2B	0.001	0.028	<u>0.102</u>	0.026	<u>0.103</u>	0.046
	2C	-0.009	0.028	<u>0.096</u>	0.025	0.087	0.044
	2D	-0.007	0.028	<u>0.094</u>	0.026	0.086	0.045
	2E	-0.013	0.027	<u>0.093</u>	0.024	0.080	0.043
	2F	-0.006	0.027	<u>0.100</u>	0.024	<u>0.094</u>	0.043

See notes on next page.

Notes to Table 3:

- (1) Models 1 and 1A-1F have equal error variances for all vehicle types. Models 2 and 2A-2F have different error variances for each vehicle type.
- (2) Models 1 and 2 are the final models based on all the data except for two outlier phase averages F150-4-0 and LeSabre-1-0.  
Models 1A and 2A are the same models fitted to all the data.  
Models 1B and 2B are the same models fitted to all the data except for the two outlier vehicles F150-4 and LeSabre-1.  
Models 1C and 2C are the same models fitted to all the data except for the outlier phase average F150-4-0.  
Models 1D and 2D are the same models fitted to all the data except for the outlier vehicle F150-4.  
Models 1E and 2E are the same models fitted to all the data except for the outlier phase average LeSabre-1-0.  
Models 1F and 2F are the same models fitted to all the data except for the outlier vehicle LeSabre-1.
- (3) Statistically significant effects at the five percent level are underlined.

## REFERENCES

- Barnett, V., and T. Lewis. 1994. *Outliers in Statistical Data*. Wiley, New York.
- Cohen, J. P., and A. M. Noda. 1995. *Auto/Oil Air Quality Improvement Research Program: Description of Phase II Working Data Set*. Systems Applications International (SYSAPP-95/048).
- Rosner, B. 1975. On the detection of many outliers. *Technometrics*, 17:221-227.